JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 1223-1235 (2018)
DOI: 10.6688/JISE.201809_34(5).0007

Predicting Student Performance in MOOCs Using Learning Activity Data

YU-CHEN CHIU[1], HWAI-JUNG HSU[1], JUNGPIN WU[2] AND DON-LIN YANG[1,+]
[1]*Department of Information Engineering and Computer Science*
[2]*Department of Statistics*
Feng Chia University
Taichung, 407 Taiwan
E-mail: {M0407060; hjhsu; cwu; dlyang}@fcu.edu.tw

Massive Open Online Courses (MOOCs) allow students to study anytime, anywhere via the internet. Unfortunately, low completion rates and the enormous number of students make it difficult for instructors to monitor student progress. Students who perform poorly are susceptible to giving up due to a lack of appropriate counseling. Increasing the number of teaching assistants may ease the situation; however, this can be prohibitively expensive.

In this study, we sought to improve the completion rate of MOOCs by predicting student performance through the analysis of data related to learning behavior and intervening before a student gives up. Learning behavior was first collected from OpenEdu, a well-known MOOC platform in Taiwan. A statistical model was then used to predict the performance of students based on this data. The effectiveness of the proposed model was demonstrated using cross validation of data from actual courses. Our findings demonstrate the effectiveness of monitoring student behavior in answering questions, watching videos, and participating in discussions on forums with the aim of predicting student performance in MOOCs in order to improve completion rates.

Keywords: MOOCs, data mining, machine learning, linear regression, logistic regression, software model generation

1. INTRODUCTION

Massive Open Online Courses (MOOCs) [1] allow people to take courses from a variety of institutions and instructors via the internet anytime, anywhere. Despite the growing popularity of systems such as Coursera [2], Open edX [3], and Udacity [4], they are also plagued with the issue of low completion [5]. The fact that the number of students in a MOOC can exceed by several times the enrollment in conventional courses makes it difficult for instructors to keep track of the learning situation of students. Many students are unable to keep up with the course and students performing poorly in any stage of a course are prone to giving up due to a lack of appropriate counseling. Assisting students who are performing poorly or lacking in motivation is essential to the success of the platform [6]. Increasing the number of teaching assistants can make a difference; however, the cost of this can be prohibitive. Identifying students who need help and providing help when it is most needed would be a more efficient approach to counseling.

Analyzing data from learning management systems (LMSs) has proven effective in improving the quality of courses and facilitating assistance for students [7, 8]. The fact

Received September 16, 2017; revised December 4, 2017; accepted March 14, 2018.
Communicated by Chang-Shing Lee.

that MOOC platforms record all of the activities of participants makes it relatively easy to perform fine-grained analysis of user behavior. In [9, 10], the researchers examine the behavior of users navigating web-pages and course videos to evaluate MOOC platforms.

In this study, we sought to predict learning performance using data pertaining to learning behavior in an online environment. Learning behavior was first collected from OpenEdu [11], a well-known MOOC platform in Taiwan. We then analyzed thirteen features of student activity to establish a statistical model to predict learning performance. In other words, we sought to elucidate the relationship between learning behavior and learning performance. Our objective was to notify instructors of situations requiring intervention in order to improve the completion rate of MOOCs.

2. RELATED WORK

2.1 Challenges in Analysis of MOOC Data

The amount and depth of data collected by MOOC platforms is growing rapidly. The millions of page clicks generated in a typical online course contain structured and unstructured data as well as spatial and temporal information. Extracting useful information from this data requires engineering methods to transform MOOC data into analyzable features. Most MOOC administrators are unfamiliar with data mining and analytics techniques. It is therefore important to provide intuitive and easy-to-understand results from the analysis of MOOC data.

2.2 Regression Analysis

Regression analysis is a statistical method [12] aimed at elucidating the causal relationship between two or more variables based on mathematical models. For example, this method can be used to predict future trends in economic growth based on prices, demand for stock, and the supply of funds. Regression analysis can be divided into simple linear regression, multivariate regression, and logistic regression.

Linear regression is used to find the linear function that best fits the relationship between a two-dimensional vector of one independent variable and one dependent variable. The aim is to predict the dependent variable using the independent variable. Multivariate linear regression is an extension of linear regression using multiple independent variables [13]. Logistic regression [14] is applied in cases where the dependent variable is binary, such as the relationship between sugar intake and diabetes mellitus. These models are used to predict the binary dependent variable based on one or more independent variables that do not necessarily conform to the normal distribution. Logistic regression is widely used in sociology, biostatistics, and marketing.

2.3 Correlation Analysis

Correlation [15] is used to illustrate the strength and direction of the linear relationship between two random variables. For example, when the national economic performance and birth rates increase simultaneously, the two variables are said to be positively correlated. Conversely, when an increase in prices corresponds to a drop in demands, we say that the two variables are negatively correlated.

The Pearson correlation coefficient [16] is a well-known metric used in correlation analysis. A coefficient close to 1 indicates a strong positive correlation, whereas a coefficient approaching −1 indicates a strong negative correlation. A coefficient close to 0 indicates that no linear relationship exists between the two variables.

3. METHODS

In this section, we outline the data collected from OpenEdu and the methods adopted for data extraction, preprocessing, correlation analysis, and model establishment.

3.1 Overview of MOOC Platform Data

OpenEdu is an MOOC platform based on Open edX [3], which provides learning activity log files as well as detailed explanations of each column in the log. Fig. 1 presents the data repository architecture of the OpenEdu platform. The data repository comprises a MySQL database, a MongoDB database, and a Tracking Log repository. The MySQL database stores the user profile, enrollment course records, and course information. The MongoDB database stores data pertaining to forums, videos, and exercises. The Tracking Log provides a record of user activity on the website, including video playback, forum discussions, performing exercises, and page browsing. All data are time stamped and stored in JSON format on the server side.

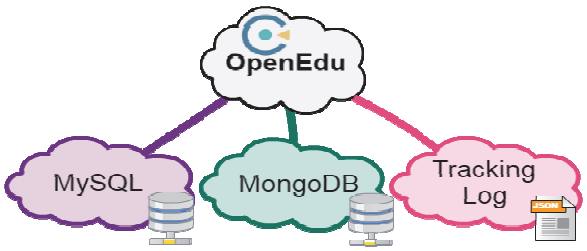

Fig. 1. Data available from OpenEdu.

3.2 Data Preprocessing

Fig. 2 presents the steps of data preprocessing. Data rows with empty and/or garbled values are first removed from the Tracking Log in the Data Cleaning process. Tripartite data from filtered Tracking Log, MySQL, and MongoDB are then preprocessed using *R* to create tables listing student and course information. The data is normalized after preprocessing to eliminate differences among the various courses.

In this study, we selected nine courses from a variety of domains. All courses included at least 100 students. A total of 5,537 students were included in the analysis. Student learning activity was isolated by removing all data pertaining to teachers and staff. During model building, we compensated for the low completion rate of OpenEdu courses by filtering out students with final grades of 0. This left a total of 1,313 valid users for the final model building. Table 1 presents an overview of the nine courses,

where "N" refers to the number of the users taking the courses and "Valid" refers to the number of valid users after filtering.

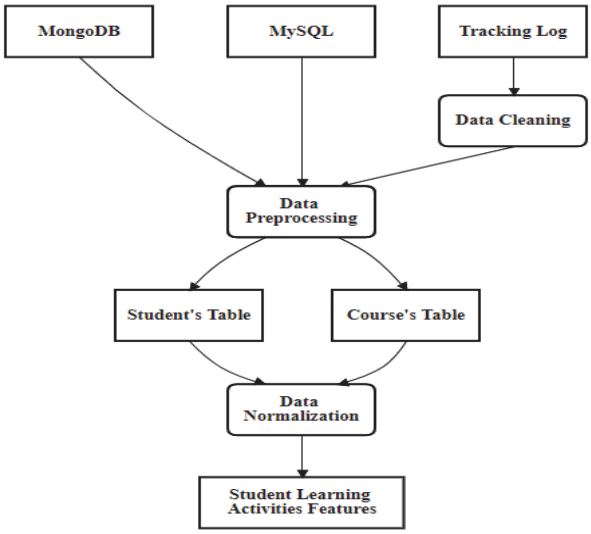

Fig. 2. Data preprocessing method.

Table 1. Overview of nine courses analyzed in this study.

	Course Name	Domain	N	Valid
C1	Microprocessor embedded system	Information engineering	211	26
C2	Entrepreneurship and business growth	Business management	1,028	183
C3	Project management	Business management	1,316	250
C4	Spatial information	Applied science	368	21
C5	Acoustics	Applied Science	405	35
C6	Physics-1	Natural science	1,301	345
C7	Taiwan Hakka culture	social science	103	61
C8	Physics-2	Natural science	362	208
C9	Make MOOCs	social science	443	184
			5,537	1,313

To formulate an accurate prediction of learning performance, we extracted the features of learning activity for model building based on the following considerations:

(1) Learning motivation
We used the time spent on a course (*e.g.*, the time spent watching videos and the number of days/weeks in the course) as a metric of learning motivation and efforts expended by students.

(2) Course material usage
Students who are more committed to the course access course materials more often than do students who are less committed. Thus, we used the number of videos

watched by students and the total number of viewings (including repeated views) as an indicator of course material usage.

(3) Course engagement

Students who are actively engaged in a course are more likely to participate in forums. Thus, we adopted the number of posts and responses on the corresponding forum as an indicator of engagement in the course.

(4) Completed exercises

The completion of exercises is essential to one's achievement in a course. Thus, we adopted the number of exercises taken by the student as an indication of expended effort.

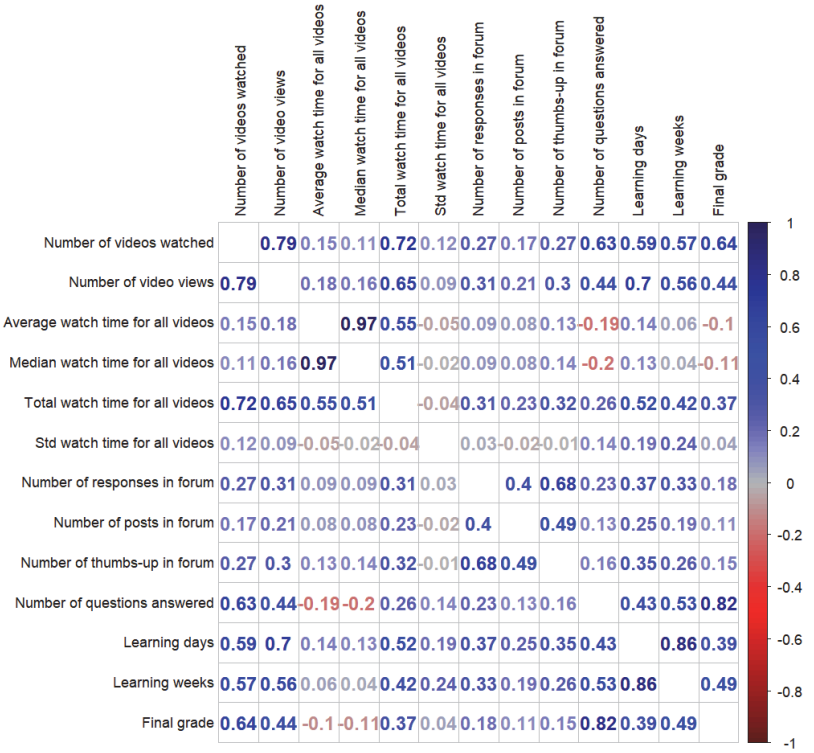

Fig. 3. Pearson's correlation coefficient of features pertaining to student behavior.

Based on the above, we derived thirteen features for each valid user in each of the nine courses. The features were as follows: *Number of videos watched*, *Number of video views*, *Average/Median/Total watch time for all videos* (*i.e.*, Average/Median/Total time spent watching videos), *Std watch time for all videos* (*i.e.*, Standard deviation of time spent watching videos), *Number of responses in forum*, *Number of posts in forum*, *Number of thumbs-up in forum*, *Number of questions answered*, *Learning days/weeks* (*i.e.*, Number of days/weeks participating in course), and *Final Grade*. Final Grade was used as the sole metric of learning performance, whereas the other features were used as indicators of learning behavior.

We used the Pearson correlation coefficient for each pair of features to analyze the relationships among the various features, the results of which are presented in Fig. 3.

Correlation analysis revealed several pairs of features that were highly correlated:

(1) *Average watch time for all videos* ↔ *Median watch time for all videos*
(2) *Learning days* ↔ *Learning weeks*
(3) *Number of questions answered* ↔ *Final grade*

All of the correlation coefficients of the above groups exceeded 0.8, which means that the features of each pair are mutually dependent. For the first two pairs, we adopt the findCorrelation function in the software package Caret to identify features that are mutually correlated. The feature with the lowest correlation was retained as representative of mutually correlated features. Therefore, *Median watch time for all videos* and *Learning weeks* were retained for modeling, whereas *Average watch time for all videos* and *Learning days* were removed to avoid the problem of collinearity.

We conducted a detailed examination of the correlation between the *Number of questions answered* and *Final grade*. Table 2 lists the correlation coefficients between the *Number of questions answered* and *Final grade* for each and all of the nine courses (C1-C9). All examined links presented highly correlated results with coefficients exceeding 0.8, indicating that it should be possible to evaluate the performance of students based on the effort they expend in doing exercises.

Table 2. Correlation between *Number of questions answered* and *Final grade.*

	Course									
	all	C1	C2	C3	C4	C5	C6	C7	C8	C9
Number of questions answered	0.82***	0.99***	1***	0.97***	0.82***	0.99***	0.89***	0.98***	0.95***	0.95***
N	1313	26	183	250	21	35	345	61	208	184
*p<.0.05, **p<0.01, ***p<0.001										

3.4 Establishing Model to Predict Learning Performance Based on Learning Activities

All features were normalized using course information to ensure the applicability of the proposed model across different courses. For example, the total numbers of videos viewed by participants varied among the nine courses. We therefore divided the *Number of videos watched* by all students in the nine courses by the number of videos in the corresponding courses. We also converted the student's *Final grade* into a binary value of 1 (pass) or 0 (fail) in accordance with the thresholds of the individual courses.

Fig. 4 presents the processing involved in predicting and validating learning performance following feature normalization. The data set was divided into training data and validation data. We adopted both linear and logistic regression models according to our purpose. Linear regression was used to observe differences in the impact of learning activities on learning performance. Logistic regression was subsequently used to predict whether a student would pass that course. In this manner, we were able to identify the students who require counseling as well as the reasons they need counseling.

Our results were verified using 10-fold cross-validation. The data set was divided into ten parts, each of which was used discretely for validation while the other nine parts were used for model training. When applying the linear regression model, we recorded

the *R*-square result following each validation and took the average of all *R*-square values as the validation result. When applying the logistic regression model, we generated a confusion matrix with each validation to calculate the accuracy, precision, and recall values of the logistic regression model.

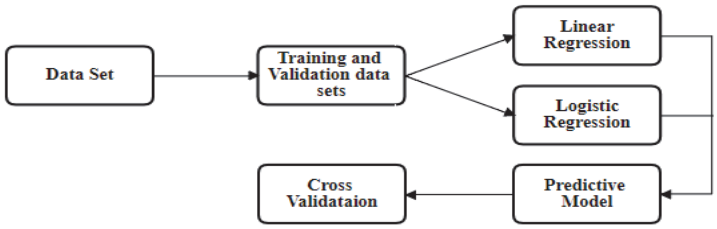

Fig. 4. Processes involved in predicting and validating learning performance.

4. PREDICTION RESULTS

4.1 Regression Model for Prediction of Learning Performance

After removing the correlated features in correlation analysis, the remaining ten features, *Number of videos watched*, *Number of video views*, *Median watch time for all videos*, *Total watch time for all videos*, *Std watch time for all videos*, *Number of responses in forum*, *Number of posts in forum*, *Number of thumbs-up in forum*, *Number of questions answered*, and *Learning weeks* are adopted as the independent variables to predict whether a student passes a course. As mentioned in the previous section, whether a student passes a course is decided by the student's *Final Grade* and the passing threshold of the course. Therefore, a student's *final grade* of a course is converted to a binary value of 1 (pass) or 0 (fail). Then, a logistic regression model to predict a student's learning performance from learning behavior can be established.

To confirm the predictability of learning performance in different courses, we use 10-fold cross-validation to avoid the problem of overfitting. Table 3 shows the model of learning performance prediction using logistic regression. The accuracy of the model is 93.5%. The precision of the model is 97.7%. The recall of the model is 89.7%. The model has a high predictability for learning performance. We use the linear regression model to discuss the impact of learning performance from learning activities.

We trained several linear regression models and analyzed the difference for all courses and each individual course. The final grade is used as the dependent variable and learning activity features as the independent variables. Table 4 shows the results of learning performance prediction using linear regression. "all" shows the result of analyzing all nine courses together while C1 to C9 represent nine individual courses respectively. We find that in all models, *Number of questions answered* is significant with a high regression coefficient. The models' *R*-squares are between 0.78 and 0.99. Therefore, we have confirmed the high correlation between the number of questions answered and learning performance as expected.

Table 3. Learning performance prediction model based on logistic regression.

	Fold										
Features	fold_1	fold_2	fold_3	fold_4	fold_5	fold_6	fold_7	fold_8	fold_9	fold_10	
(Intercept)	-10.26***	-9.30***	-11.7***	-9.84***	-10.3***	-9.7***	-10.0***	-9.67***	-9.72***	-9.9***	
Number of viedeos watched	2.35**	2.43**	2.7**	2.24**	2.5**	2.5**	2.4**	2.53**	2.37**	2.5**	
Number of viedeo views	-1.39***	-1.29***	-1.4***	-1.17***	-1.2***	-1.4***	-1.3***	-1.41***	-1.24***	-1.5***	
Median watch time for all videos	-0.32	-0.3	-0.3	-0.16	-0.2	-0.2	-0.4	-0.3	-0.33	-0.2	
Total watch time for all videos	-0.03	0.07	-0.4	-0.44	-0.4	-0.2	-0.3	-0.08	-0.39	-0.2	
Std watch time for all videos	0.46	-0.02	0.4	0.09	0.2	0.3	0.5	0.21	0.08	-0.2	
Number of responses in forum	-3.41**	-3.91***	-3.9***	-3.62***	-3.7***	-3.8***	-4.7***	-3.76***	-3.69***	-3.8***	
Number of posts in forum	-1.48	-0.72	-1.6.	-0.44	-1.4	-1.1	0.1	-0.77	-0.82	-0.8	
Number of thumbs-up in forum	10.48***	9.08***	12.1***	9.85***	10.6***	9.9***	10.4***	9.83***	9.99***	10.1***	
Learning weeks	1.65*	1.07	1.3.	1.28.	1.3.	1	1	1.06	1.23.	1.1	
Number of questions answered	11.78***	11.02***	13.3***	11.52***	11.8***	11.4***	11.6***	11.29***	11.43***	11.8***	Mean
Accuracy	0.95	0.92	0.92	0.93	0.93	0.96	0.94	0.94	0.93	0.93	0.935
Precision	0.98	1	0.97	1	0.95	1	0.98	0.97	0.97	0.95	0.977
Recall	0.91	0.85	0.88	0.87	0.91	0.93	0.9	0.91	0.9	0.91	0.897
N	1313	1313	1313	1313	1313	1313	1313	1313	1313	1313	
. p<0.1, *p<.0.05, **p<0.01, ***p<0.001											

Table 4. Learning performance prediction model based on linear regression.

	Course									
Features	all	C1	C2	C3	C4	C5	C6	C7	C8	C9
(Intercept)	-0.045*	0.006	0.0028	-0.1661***	-0.037	-0.01	-0.063.	-0.009	-0.1670***	-0.006
Number of viedeos watched	0.174***	-0.104	-0.0091	-0.0112	-0.332	-0.051	-0.320*	0.055	0.0102	-0.177***
Number of viedeo views	-0.050***	0.049	0.0155.	-0.0134	0.308	0.008	0.181	-0.021*	0.0792	0.036**
Median watch time for all videos	-0.02	-0.075	0.0009	0.0171	-0.189	-0.012	0.008	-0.054.	0.0127	-0.006
Total watch time for all videos	-0.001*	0.001	-0.0001.	0.0006	0.008	0.003	0.001	0.005*	-0.0008	0.001
Std watch time for all videos	-0.082*	-0.019	-0.0108.	-0.0188	0.125	0.001	-0.116.	0.021	-0.0026	0.083**
Number of responses in forum	-0.364***	0.144	0.0129	0.0095	0.292	-0.047	0.172	-0.079.	0.1055	0.309***
Number of posts in forum	-0.103*	-0.049	0.0587**	0.0367	-0.309	0.066	-0.128	0.018	0.0245	0.226**
Number of thumbs-up in forum	0.538***	-0.046	0.0025	NA	-0.061	0.037	0.066	NA	-0.1343	-0.042
Learning weeks	0.103**	-0.002	-0.0307*	-0.0286	0.2	0.043	0.124*	0.083*	0.0694	-0.017
Number of questions answered	0.822***	1.013***	0.6950***	2.2961***	0.741***	0.569***	0.975***	0.873***	0.9543***	0.654***
R^2	0.78	0.99	0.99	0.94	0.86	0.98	0.8	0.96	0.9	0.94
N	1313	26	183	250	21	35	345	61	208	184
10-folds validation R^2	0.78									
. p<0.1, *p<.0.05, **p<0.01, ***p<0.001										

4.2 Regression Model Excluding Number of Questions Answered

We observed a strong correlation between the number of questions answered and learning performance; however, we also sought to identify other features that could have a significant impact on learning performance. An awareness of what learning activities are most effective could help instructors to provide more meaningful guidance. We generated additional logistic and linear regression models without the feature of *Number of questions answered*. Table 5 presents the logistic regression model without the number of questions answered. We observed that the overall accuracy of this model decreased from 93.5% to 78.9%, indicating that the *Number of questions answered* does indeed have considerable influence on predicting learning performance. Nonetheless, the other

learning behaviors (*Number of videos watched*, *Number of thumb-ups in forum*, *and Learning weeks*) still present a high degree of reliability in predicting learning performance.

Table 5. Logistic regression model excluding *Number of questions answered.*

	Fold										
Features	fold_1	fold_2	fold_3	fold_4	fold_5	fold_6	fold_7	fold_8	fold_9	fold_10	
(Intercept)	-2.8***	-2.62***	-2.7***	-2.76***	-2.7***	-2.8***	-2.624***	-2.9***	-2.7***	-2.7***	
Number of viedeos watched	7.8***	7.53***	7.8***	7.93***	7.8***	8.0***	7.880***	8.8***	7.8***	7.9***	
Number of viedeo views	-1.4***	-1.46***	-1.5***	-1.52***	-1.4***	-1.6***	-1.373***	-2.0***	-1.4***	-1.6***	
Median watch time for all videos	-0.6*	-0.77**	-0.8**	-0.75*	-0.7*	-0.8**	-0.777**	-0.5	-0.8**	-0.6*	
Total watch time for all videos	-3.1***	-2.52***	-2.7***	-2.75***	-3.0***	-2.7***	-3.005***	-3.3***	-2.9***	-3.0***	
Std watch time for all videos	-1.1*	-1.55**	-1.1*	-1.09.	-1.4*	-0.9.	-1.669**	-1.5*	-1.4*	-1.4*	
Number of responses in forum	0.5	-0.15	0.2	-0.07	0.4	1.3	0.445	0.3	0.4	0.7	
Number of posts in forum	0.1	0.05	-0.2	0.47	0.2	-0.3	-0.009	0.8	-0.2	-0.3	
Number of thumbs-up in forum	2.0***	2.00***	1.9***	2.07***	2.0***	1.9***	1.765***	2.1***	2.0***	2.0***	
Learning weeks	3.1***	3.23***	3.1***	3.00***	3.1***	3.2***	3.150***	3.4***	3.2***	3.2***	Mean
Accuracy	0.81	0.8	0.8	0.79	0.8	0.75	0.77	0.76	0.77	0.84	0.789
Precision	0.81	0.76	0.83	0.77	0.8	0.77	0.78	0.8	0.78	0.84	0.794
Recall	0.82	0.9	0.76	0.82	0.82	0.71	0.76	0.71	0.78	0.85	0.793
N	1313	1313	1313	1313	1313	1313	1313	1313	1313	1313	
. p<0.1, *p<.0.05, **p<0.01, ***p<0.001											

Table 6. Linear regression model excluding *Number of questions answered.*

	Course									
Features	all	C1	C2	C3	C4	C5	C6	C7	C8	C9
(Intercept)	0.245***	0.72**	0.019	0.53***	0.53**	0.08	0.165**	0.091.	-0.07	0.005
Number of viedeos watched	0.826***	-0.23	1.538***	0.50**	-0.48	-0.02	1.287***	0.835***	1.45***	0.104*
Number of viedeo views	-0.146***	-0.25	-0.549***	-0.07	0.47	0.07	-0.232	-0.039*	-0.32*	0.089***
Median watch time for all videos	-0.297***	-0.44	0.017	-0.40***	-0.44	-0.06	-0.054	-0.105*	0.04	-0.033
Total watch time for all videos	0.001	0.02	-0.005***	0.01*	0.05	0.02***	-0.012**	0.005	-0.02**	0.006.
Std watch time for all videos	-0.144**	-0.87	-0.045	0.18	0.21	-0.06	-0.364***	0.219.	0.04	0.002
Number of responses in forum	-0.113.	0.05	0.003	0.24	0.42	0.14	-0.009	-0.113	0.38**	0.538***
Number of posts in forum	-0.04	0.36	0.058	0.03	-0.54	-0.12	-0.044	0.023	0.30*	0.499***
Number of thumbs-up in forum	0.258***	0.15	-0.052	NA	-0.08	0.11	0.061	NA	-0.24	-0.079
Learning weeks	0.448***	0.87	0.700***	0.07	-0.11	-0.08	0.485***	0.115.	0.45***	-0.06
R^2	0.51	0.65	0.79	0.32	0.5	0.85	0.52	0.87	0.67	0.85
N	1313	26	183	250	21	35	345	61	208	184
10-folds validation R^2	0.49									
. p<0.1, *p<.0.05, **p<0.01, ***p<0.001										

Table 6 presents the linear regression model in which the feature of *Number of questions answered* was excluded. In the case of a single course, the coefficients of *Number of videos watched* had a significant effect in the models of all courses with a class size greater than fifty (six out of the nine courses). We also observed a correlation between the time spent studying course materials and learning performance. The number of weeks participating in course (*Learning weeks*) had a significant influence in four models, which suggests that persistence can affect learning performance. In Table 6,

courses C8 and C9 both generated intensive discussions in corresponding forums, such that the *Number of responses in forum* and *Number of posts in forum* having significant effects. Our results indicate that participation in online discussions is strongly correlated with learning performance in some courses.

In the above analysis, we learned that the impact of specific learning activities varies among courses. Nonetheless, the logistic regression model trained using learning activity features without *Number of questions answered* is predictive of the learning performance of individual students in all of the courses. This is a clear demonstration of the effectiveness of these features in predicting learning performance.

4.3 Regression Model for Predicting Learning Performance on a Weekly Basis

The models presented in Section 4.2 are meant to predict learning performance based on data covering the duration of a course. However, improving the course completion rate requires that instructors identify the students who need help as early as possible in order to provide counseling before the student gives up. In other words, it is necessary to predict the final performance of students as early as possible. To achieve this, we established a logistic regression model based on learning activity on a week-by-week basis. Our aim was to allow the prediction of final performance in a course through analysis of weekly learning behavior data.

In this section, we employed cumulative data obtained on a weekly basis to predict final learning performance in two Physics courses. The two courses included the same content and ran for the same duration, thereby allowing the combination of results in order to acquire sufficient data for model training.

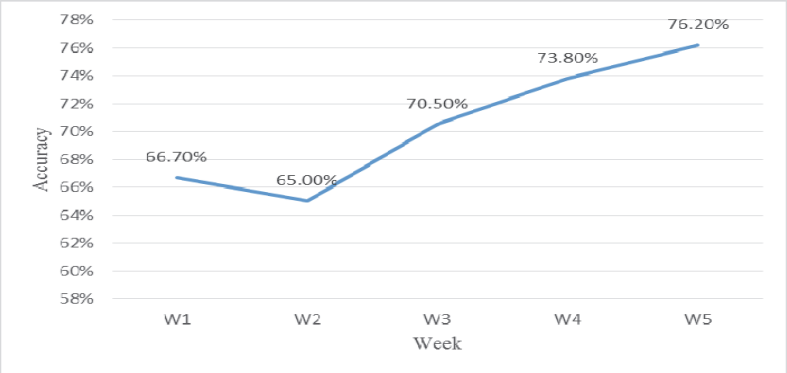

Fig. 5. Accuracy of logistic regression model of Physics courses on a weekly basis.

Fig. 5 shows the accuracy of predictions pertaining to whether a student would pass the Physics course using weekly cumulative data of learning behavior. The duration of the course was six weeks. Data from the first to the fifth weeks was used to train the logistic regression model aimed at predicting the final performance of students taking the course. We used 10-fold cross-validation using data obtained on a week-by-week basis to evaluate the accuracy of predicting whether a student would pass the Physics course. As shown in Fig. 5, the accuracy of predicting learning performance in the first two

weeks was as follows: week 1 (66.7%), weeks 1 and 2 (65%). Beyond the first two weeks, the accuracy gradually increased, as follows: weeks 1-3 (70.5%), weeks 1-4 (73.8%), and weeks 1-5 (76.2%). These results suggest the possibility of forecasting final performance in an online course by monitoring learning behavior on a weekly basis. The MOOC platform could be configured to automatically remind instructors of the need to offer counseling to students having difficulties following the course.

5. CONCLUSIONS AND FUTURE WORK

In this study, we established a data-driven approach to the analysis of learning behavior data to predict learning performance on an MOOC platform. Four types of learning activity were adopted: (1) learning motivation; (2) course material usage; (3) course engagement; and (4) completed exercises. This system could be used to assist instructors in tracking the learning performance of students in MOOCs to identify students requiring counseling as early as possible, and thereby improve the completion rate.

We established a logistic regression model to predict whether a student would pass a course as well as a linear regression model to predict the student's final grade in a given course. The logistic regression model could be used to identify students requiring assistance, whereas the linear regression model could be used to identify the factors influencing student performance in various courses.

Correlation analysis and observations of the coefficients of the prediction model revealed that *Number of questions answered* is highly correlated with *Final grade.* The fact that the model with *Number of questions answered* achieved accuracy of 93.5% indicates that it could be used as the primary metric by which to evaluate learning performance. When the models were retrained without *Number of questions answered*, the accuracy dropped to 78.9%, which indicates that using the remaining features to predict the final performance of students in online courses is still feasible. Improving the completion rate of MOOCs requires that students in need of assistance be identified as early as possible. We therefore established another model to predict final performance based on cumulative data obtained weekly from the first week to the last week. Our results show that the monitoring of learning behavior in the middle of the course can be used to predict whether a student will pass (with accuracy exceeding 70%). This should provide sufficient time for intervention.

In the future, we will examine the relationship between learning behavior and learning performance from other theoretical perspectives with the aim of improving the accuracy of predictions. The current results were obtained using the Open edX-based platform; however, the proposed system is also applicable to other MOOC platforms.

ACKNOWLEDGMENT

This research was supported in part by the Ministry of Science and Technology, Taiwan, under grant numbers MOST 105-2634-E-035-001, 105-2221-E-035-082 and 106-2221-E-035-097.

REFERENCES

1. J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld, "MOOCs: So many learners, so much potential," *IEEE Intelligent Systems*, Vol. 28, 2013, pp. 70-77.
2. C. Severance, "Teaching the world: Daphne Koller and coursera," *Computer*, Vol. 45, 2012, pp. 8-9.
3. L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom: Research into edX's first MOOC," *Research & Practice in Assessment*, Vol. 8, 2013, pp. 13-25.
4. Udacity, https://en.wikipedia.org/wiki/Udacity/, June, 2017.
5. J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction," in *Proceedings of the 11th International Conference on Computer Science & Education*, 2016, pp. 52-57.
6. A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Computer*, Vol. 49, 2016, pp. 61-69.
7. R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Transactions on Learning Technologies*, Vol. 10, 2017, pp. 17-29.
8. J. K. Tarus, Z. Niu, and A. Yousif, "A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining," *Future Generation Computer Systems*, Vol. 72, 2017, pp. 37-48.
9. C. G. Brinton and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *Proceedings of IEEE Conference on Computer Communications*, 2015, pp. 2299-2307.
10. C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from massive open online courses," in *Proceedings of IEEE Pacific Visualization Symposium*, 2015, pp. 159-166.
11. OpenEdu, https://copeneduc.org/, June, 2017.
12. D. M. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, *Introduction to Statistics*, Rice University, Houston, Texas, 2013.
13. D. A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, NY, 2009.
14. D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 20, 1958, pp. 215-242.
15. C. F. Dietrich, *Uncertainty*, *Calibration and Probability: the Statistics of Scientific and Industrial Measurement*, CRC Press, Florida, 1991.
16. S. M. Stigler, "Francis Galton's account of the invention of correlation," *Statistical Science*, Vol. 4, 1989, pp. 73-79.

Yu-Chen Chiu (邱毓宸) received his M.S. degree in Computer Science from the Department of Information Engineering and Computer Science, Feng Chia University, Taiwan, in 2017. His main research interests include data mining, machine learning, and software engineering.

Hwai-Jung Hsu (許懷中) received his B.E., M.S., and Ph.D. degrees in Computer Science from National Chiao Tung University, Taiwan in 2001, 2003, and 2011 respectively. He worked at Institute of Information Science, Academia Sinica from 2011 to 2016 as a postdoctoral researcher. In 2016, he joined the faculty of Feng Chia University as an Assistant Professor in the Department of Information Engineering and Computer Science. His research interests include big data analytics and applications, software engineering, cloud computing, psychophysiology and crowdsourcing.

Jungpin Wu (吳榮彬) received the B.S. degree in Applied Mathematics from Tatung University, Taiwan in 1988, the M.S. degree in Statistics from the Graduate Institute of Statistics of National Central University, Taiwan in 1993, and the Ph.D. degree in Statistics from the North Carolina State University in 1998. He was a postdoctoral staff at Academia Sinica from 1998 to 1999. Since then, he joined the faculty of Feng Chia University and is currently an Associate Professor in the Department of Statistics. His research interests include spatial statistics, generalized estimateing equations, empirical process approach, and data mining.

Don-Lin Yang (楊東麟) received the B.E. degree in Computer Science from Feng Chia University, Taiwan, in 1973, the M.S. degree in Applied Science from the College of William and Mary in 1979, and the Ph.D. degree in Computer Science from the University of Virginia in 1985. He worked at IBM Santa Teresa Laboratory from 1985 to 1987 and at AT&T Bell Laboratories from 1987 to 1991. Since then, he joined the faculty of Feng Chia University and is currently a Professor in the Department of Information Engineering and Computer Science. He received 2015 IEET Distinguished Teaching Award. His research interests include data mining, database system, and software engineering. Dr. Yang is a member of the IEEE Computer Society and the ACM.